Process for encoding speech and terminals for implementing the process.

The transmission of speech on the switched telephone network STN necessitates a pass-band sufficient for the speech to remain comprehensible. A band ranging from very low frequencies to some kilohertz represents a good compromise between fidelity of restoration and pass-band resources. For this reason, in order to transmit the voice on the STN, of which the inter-centre connections are digital, the voice frequencies are encoded to transform them into a digital signal at the basic rate of the STN, representing the evolution of the amplitude of the voice signal.

However, it is sometimes desirable to reduce the rate of the transmission, for example, in a voice synthesis terminal, of which the message memory must remain of limited size. In the same way, it may be desired to use only one transmission channel of a specific network, with an output well below the afore-mentioned rate of the telephone network.

In such a case, an attempt is made to recognise the various phonemes of a voice sequence instantly. These phonemes are compared with references in a library, which are associated with code words and these phonemes are replaced by the corresponding code words which describe the speech using a much smaller quantity of information. In this way the voice is compressed.

During reception, the called terminal comprises the same library and, by means of voice synthesis, restores analogue signals corresponding to the various code words.

However, such a procedure presents the disadvantage of restoring only a voice which has been standardised by the library and which is thus impersonal, and it is therefore essentially impossible to recognise the correspondent in order to authenticate a voice message. The inflexions or fluctuations of the voice, which form part of the information just as much as the meaning of the words themselves, are thus not restored.

The present invention aims to achieve voice encoding which makes it possible both to compress the information and to achieve a personalised restoration.

To this end, the invention firstly relates to a process for encoding speech formed from a sequence of acoustic units, in which the units are compared with library references associated with primary code words, the differences between the units and the references are determined,

New page 1a filed in response to the first written opinion

The article .Speaker normalization algorithms for very low-rate speech coding. of Roucos et al, ICASSP 84, Vol. 1, 19-21 March 1984, page 1.11 to 1.14, XP002073267, SAN DIEGO, CA, US, and the patent DE-A-3 416 238 disclose the progressive adaptation of a library of voice signals to those of a new speaker. The differences found between these signals serve only for the purposes of this adaptation.

The speech signals are thus deformed upon restoration while the library has not reached a new state of equilibrium.

WO-A-94 18 668 relates to the transmission of speech by linear prediction and does not relate to voice recognition.

Patent Abstracts of Japan, Vol. 010, No. 240, JP 61 071 730A discloses recognition, apart from the speech, of a voice characteristic but without indicating either its nature or, in particular, the manner of determining this characteristic.

Faithful reproduction of the speech cannot be ensured.

the differences are encoded by secondary code words and pairs of primary and secondary codes are substituted for the units.

Thus the primary code words will effectively and compactly encode the largest part of the acoustic energy input, while the secondary code words will improve the fidelity of restoration but without greatly increasing the volume of code data since they relate to only a limited energy and a low number of bits makes it possible to encode this marginal energy modulating the primary, standard energy corresponding to the primary code words.

The invention also relates to a terminal for encoding speech signals, comprising means for inputting a sequence of acoustic units and transmitting it to comparator means arranged to compare the acoustic units successively with library references and thus select therein in each instance a specific primary code word of one of the references, the terminal being characterised by the fact that the comparator means are arranged to determine a difference between the input acoustic unit considered and the reference corresponding to the code word selected and to transmit this difference to transcoding means provided to supply, in response, a secondary code word corresponding to memory means arranged to associate the respective primary and secondary code words.

Finally, the invention relates to a terminal for decoding signals, comprising means for receiving signals representing primary code words of references of acoustic units in a library, and decoding means arranged to select certain ones of the references in the library according to the primary code words received and to control a transducer for restoration of speech signals accordingly, the terminal being characterised in that the decoding means are arranged also to decode secondary, correction code words associated with the primary code words, and to correct the selected voice references accordingly.

Although the process of the invention makes it necessary, all in all, to have an encoding terminal and a corresponding decoding terminal, each of these can be sold separately and the applicant thus intends to claim each one.

In particular, it is advantageous to provide a facsimile machine comprising means for inserting the code words into a facsimile message.

The invention will be better understood with the aid of the following description of a preferred embodiment of the process of the invention, with reference to the attached drawing in which:

- Figure 1 schematically illustrates a transmission terminal and a reception terminal for voice signals for implementation of the process of the invention,
- Figure 2 shows the amplitude A of a speech signal over time t,
- Figure 3 shows the amplitude K of lines of the spectrum of the speech signal 2 over the frequency F, and
- Figure 4 is a flow diagram showing the steps of the process.

The transmission terminal referenced 15 in Figure 1, which is in this case in the form of a portable handset for a radio communication network, comprises a microphone 26 for inputting the speech signal of its user, supplying an analogue/digital converter 27 connected at the output to a central microprocessor unit 28 associated with two libraries 11 and 12 of sound sequences or standardised acoustic units such as phonemes. The central unit 28 which encodes the speech controls a transmitter 29, in this case a radio transmitter, of which the transmissions are received by a reception circuit 30 of a speech restoration terminal 35. Figure 2 illustrates the amplitude of an acoustic unit according to time t and Figure 3 illustrates the amplitude K of the lines of the spectrum corresponding to a given moment.

More precisely, the central unit 28 comprises a comparator 16 to compare the acoustic units received from the converter 27 with library acoustic units. As explained in more detail with regard to Figure 4, the comparator 16 has the function of selecting the library reference which is the most similar to the signal currently analysed and also has the function of specifying this difference, ie. of providing a deviation value for each of the criteria which serve in the selection. This difference is transcoded by a transcoding circuit 17, in order to condense its expression into the form of a secondary code word which is stored in a memory 18 under the control of the comparator 16. This comparator, which has previously stored the primary code word in the memory 18, addresses and controls the writing therein so that the two code words, primary and secondary, are physically associated, as they are from a logic point of view, ie. that, for example, chaining is defined between the two memory zones containing them.

The reception terminal 35 comprises a central unit 33 carrying out inverse decoding of the speech to supply a loudspeaker 34. Two memories forming libraries 31 and 32, in this case external to the central unit 33, are connected to this central unit. The reception terminal 35 is in this case a standard terminal for receiving written messages, known as a pager, also arranged to receive voice messages. For the sake of clarity various standard dialling input, dialling transmission and data display circuits have not been shown.

The central unit 33 comprises a circuit 36 for addressing the libraries 31 and 32, the personalised library and primary library respectively, on the basis of code words received from the reception circuit 30. In response, a buffer circuit 37 receives, from the primary library 32, spectra of primary acoustic units and transmits them to a circuit 38 for modulation or composition of these spectra. The circuit 38 modulates these spectra according to the secondary code word associated with the primary code word read from the primary library 32. The circuit 38 thus combines the information of the primary and secondary code words to restore the speech signal initially captured (26). This combination can, for example, be an addition or multiplication of frequency lines followed by an inverse Fourier transformation or it can also relate directly to signal amplitudes. In this example, each type of restored acoustic unit is stored in the personalised memory 31 in order to use this latter directly if an identical pair of code words, primary and secondary, is later received. In one variation, the memory 31 could contain only modulation values which it would supply to the circuit 38 after addressing by a secondary code word.

The encoding and decoding operations will now be explained in more detail with reference to Figure 4.

In order to encode the voice a speech signal 26 is captured by the microphone 26 during a step 1 and, in this case, it is converted into a digital signal in the converter 27 in a step 2. The speech signal is then compared, in the central unit 28, with a plurality of reference signals from the library 11 in a step 3. The comparison is carried out instantly, in practice in a cyclic manner at high speed with respect to the speed of evolution of the analysed speech signal. This signal can be considered as being a sequence of acoustic units specific to a given language, such as vowels, diphthongs or hiatus, of which a representation has initially been placed in the library 11 and associated with a code word referred to as the primary code word, which is peculiar to each. When the library 11 and the libraries 12 and 32 mentioned hereinunder are being formed, a number of voice inputs from a single speaker or from several are carried out in order to form an average voice reference. However, in order to improve the efficacity of future recognition, a number of references are preferably stored (11, 12) for each acoustic unit in order tc form a range of recognition permitting differences between speakers to be accommodated.

Each acoustic unit (Figure 2) corresponds to a particular evolution of the amplitude A or energy of the speech signal and has a duration able to vary according to the talking speed of the person speaking.

The step 3 thus consists of comparing the evolution of the amplitudes of the reference signals to that of the captured signal. In order to overcome variations in the talking speed, it is possible, for example, to consider only the succession of significant modulations of amplitude (variation of energy exceeding a threshold) without associating a time value notion therewith.

In Figure 2 the vertical arrows, of which there are eight, represent the amplitude of the extremes and thus form a signature, assumed, in this case, to represent a specific acoustic unit. If, leaving the time domain of Figure 2, the frequency domain is now considered, the Fourier transform of the momentary amplitude A of the signal at any point on the curve of Figure 2 can be shown by the spectrum of frequency lines of Figure 3. In practice, it is considered that the voice energy is essentially limited to three frequency bands located respectively around 0.1 kHz and two bands between about 1 and 3 kHz as well as 5 and 7 kHz respectively.

For this reason, if the curve of the amplitudes of Figure 2 is traced over time t, the amplitude K of each line of Figure 3 will be modulated according to the evolution of the amplitude A of the speech signal.

Thus, if the succession of the spectra of Figure 3 is stored, it is possible to restore the succession of the amplitudes A of the original signal by an inverse Fourier transform.

In order to limit the number of spectra to be processed, it is possible to effect only cyclic sampling, sufficiently close not to lose information. It is also possible to be limited to a narrow set of spectra of the extremes of amplitude which are shown by the eight arrows in Figure 2. If it is desired to limit the number of spectra further, it is even possible to retain only a single spectrum representing the average of all the spectra over the considered period of time of the acoustic unit or the average of the spectra of the extremes.

As explained hereinunder in more detail, the average spectrum, or the spectra of the speech signal captured will be compared with one or more counterpart spectra of reference speech signals in the library in order, on the one hand, to select the reference speech signal (acoustic unit) which is the most similar to the captured signal and, on the other hand, to produce a signal representing the difference between the spectrum or spectra of this latter and the spectrum or spectra of the selected reference signal. The difference signal is formed into a code word, referred to as a secondary code word, and is associated with the primary code word of the selected reference signal (recognised acoustic unit) and thus constitutes an additional item of information for modulation or correction of the standardised analogue signal which will be restored from the primary code word considered.

The primary code words of the acoustic units, successively selected as the voice sequence progresses, are stored in a step 4 in order to form a message encoded according to the standard of the library 11.

Furthermore, in a step 5 some of the acoustic units captured and recognised are processed more by analysing their frequency spectrum in detail, in this case in the frequency domain by an inverse Fourier transform, as explained above, step 6.

In a step 7, the spectrum of lines j of the acoustic unit of identity i concerned or the spectra representing its evolution over time t are compared with the spectrum or spectra of the acoustic unit selected in the library 11 and which is or are contained in the associated library 12. In this way, for the or each spectrum a series of weighting coefficients Cijt (i = identity of phoneme, j = frequency rank of the line, t = time rank) is established, each indicating the amplitude or relative energy of each line j with respect to its counterpart in the library 12. In other words, these coefficients also represent, albeit indirectly, the relative difference (1 - Cijt) between the recognised acoustic unit and the corresponding reference in the library). The lines in each of the three bands actually correspond to a row of mini bands of adjacent frequencies, in which the voice energy is detected. The analysis in the frequency domain, which is selected in this case, thus provides a more detailed item of information than in the case of an analysis in the time domain of Figure 2 where only the momentary amplitude A is available.

Thus, in the case of Figures 2 and 3, the series above comprises twelve coefficients representing the twelve lines shown, so that a table of eight such series represents the acoustic unit through the eight extremes shown. Apart from the reduction of the table to a single series, it is possible to make provision to retain only a single average weighting coefficient for each of the three bands. If each coefficient is encoded on just 4 bits, the error will not exceed about 3%, which is amply sufficient to restore a voice timbre, especially since the correction signal represents little energy with respect to the normed signal which it corrects, so that the error viewed as a total is low.

It is thus possible in this case to associate with the primary code word of the acoustic unit selected, of the order of about a hundred bits (12 x 8) if each extreme is retained, or only 12 bits (4 x 3) for the three bands. As the voice timbre is particularly provided by the high frequency of the third band, it is even possible to transmit only the secondary, correction code word relating thereto.

In a step 8, the signal representing the difference in the spectra is transformed into a secondary code word representing the table or the series mentioned above. When the captured speech sequence ends, the primary code words of step 4 and the secondary code words of step 8 are associated one to one (step 9) then transmitted on a transmission network such as, for example, the switched telephone network or, in this case, a radio messaging network (step 10).

The called terminal 35 receives the message in a step 21 and, in a step 22, a primary library file 32 similar to the file of spectra 12, is read by the circuit 36 in order to extract therefrom the primary standardised spectra according to the primary code words. In a step 23 the secondary code words serve to modulate (38) the amplitudes or energies of the standardised lines read in the primary library 32 in order, in this way, to constitute the personalised library 31 of acoustic units, ie. comprising, in particular, the timbre of the captured voice. The acoustic units of the personalised library 31 are represented in digital form in the time domain after previous transformation by an inverse Fourier transform.

In a step 24, the primary code words received are read successively in order to restore the captured speech signal via the loudspeaker 34 (step 25). For this purpose the primary code words read the personalised library 31 which thus corresponds to the library 11 but which has been personalised by the features in the spectrum of the captured voice.

As stated above, the formation of the library 31 is optional and aims to store a correction for each primary code word, which avoids having to repeat the sending of the secondary code word when the same primary code word is transmitted several times. If, on the other hand, a secondary code word is transmitted systematically, this code word can evolve to follow the possible evolutions in the timbre. In this case the restored voice is both personalised and the evolution of the timbre over time is also restored.

It should also be noted that the analysis and restoration can generally relate to the whole audible frequency band from about 15 Hz to 15 kHz, even if, in practice, it can be limited to 8 kHz. The frequencies of the band from 4 to 8 kHz, cut for standard transmission by the telephone network, are in this case analysed and restored since the corresponding information is transmitted in the form of a remote command from the library 31 which already contains the lines at these frequencies, which avoids any explicit transmission thereof.

It should also be noted that if the analysis can relate to only a limited number of sufficiently characteristic frequency bands in the library 11, 12, the various signals to be restored, in library

32, contain all the lines initially input, ie. each cover, for example, a single-piece band of 15 Hz to 8 kHz.

As explained at the beginning, the invention can apply in cases not involving transmission, in order, for example to store locally a message which to be restored later, ie. this would be a tape recording function.

In another embodiment, not illustrated, the primary and secondary code words are associated with facsimile data to form a voice-data message. The message is input via the telephone usually associated with facsimile machines and is restored by the same means at the called facsimile machine. The code words emitted by a circuit such as 28 are inserted in a specific field of the message by a microprocessor managing the facsimile transmission protocol and in the same way are retrieved upon reception to be processed as explained above. It is thus possible to carry out voice annotation of a facsimile message, annotation which is transmitted, for example, as a facsimile header.